













OPEN

DATA DESCRIPTOR

# Highly multiplexed immunofluorescence images and single-cell data of immune markers in tonsil and lung cancer

Rumana Rashid <sup>1,2,3,4</sup>, Giorgio Gaglia <sup>1,2,3</sup>, Yu-An Chen <sup>2,3</sup>, Jia-Ren Lin <sup>2,3</sup>, Ziming Du <sup>1,2,3</sup>, Zoltan Maliga <sup>2,3</sup>, Denis Schapiro <sup>2,5</sup>, Clarence Yapp <sup>2</sup>, Jeremy Muhlich <sup>2</sup>, Artem Sokolov <sup>2,4</sup>, Peter Sorger <sup>2,3,6\*</sup> & Sandro Santagata <sup>1,2,3,7\*</sup>

In this data descriptor, we document a dataset of multiplexed immunofluorescence images and derived single-cell measurements of immune lineage and other markers in formaldehyde-fixed and paraffin-embedded (FFPE) human tonsil and lung cancer tissue. We used tissue cyclic immunofluorescence (t-CyCIF) to generate fluorescence images which we artifact corrected using the BaSiC tool, stitched and registered using the ASHLAR algorithm, and segmented using ilastik software and MATLAB. We extracted single-cell features from these images using HistoCAT software. The resulting dataset can be visualized using image browsers and analyzed using high-dimensional, single-cell methods. This dataset is a valuable resource for biological discovery of the immune system in normal and diseased states as well as for the development of multiplexed image analysis and viewing tools.

## Background & Summary

Tissues comprise individual cells of diverse types along with supportive membranes and structures as well as blood and lymphatic vessels. The identities, properties and spatial distributions of cells that make up tissues are still not fully known: classical histology provides excellent spatial resolution, but it typically lacks molecular details. As a result, the impact of intrinsic factors such as lineage and extrinsic factors such as the microenvironment on tissue biology in health and disease requires molecular profiling of single cells within the broader context of organized tissue architecture. Such deep spatial and molecular phenotyping is especially pertinent to the study of cancer resection tissues. These samples are routinely acquired prior to, on, and after a therapeutic intervention, providing opportunities to characterize the interplay between malignant tumor cells and surrounding immune cell populations and how those relationships are influenced over time by treatments. Understanding these relationships may elucidate biomarker signatures that predict response to therapy<sup>1,2</sup> and is particularly relevant in the case of immunotherapeutics. Many available immunotherapies, including those targeting cytotoxic T lymphocyte-associated antigen-4 (CTLA-4), programmed cell death-1 receptor (PD-1), and programmed cell death-1 ligand (PD-L1), influence interactions between tumor and immune cells to inhibit immune checkpoints and activate the immune system's surveillance of tumor cells<sup>3-7</sup>. However, even in tumor types that are highly responsive to such therapies, many patients do not benefit, and many types of tumors remain broadly refractory to these agents. A deeper understanding of immune cell states, location, interactions, and architecture (“immunophenotypes”) promises to provide new prognostic and predictive information for cancer research and treatment.

<sup>1</sup>Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States.

<sup>2</sup>Laboratory for Systems Pharmacology, Harvard Medical School, Boston, MA, United States. <sup>3</sup>Ludwig Center at Harvard, Harvard Medical School, Boston, MA, United States. <sup>4</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States. <sup>5</sup>Broad Institute of MIT and Harvard, Cambridge, MA, United States.

<sup>6</sup>Department of Systems Biology, Harvard Medical School, Boston, MA, United States. <sup>7</sup>Department of Oncologic Pathology, Dana Farber Cancer Institute, Boston, MA, United States. \*email: [peter\\_sorger@hms.harvard.edu](mailto:peter_sorger@hms.harvard.edu); [ssantagata@bics.bwh.harvard.edu](mailto:ssantagata@bics.bwh.harvard.edu)

With recent advances in multiplexed imaging technologies<sup>8</sup>, multiple epitopes can be detected within a tissue section and the spatial distributions and interactions of cell populations precisely mapped. One such method is tissue-based cyclic immunofluorescence (t-CyCIF)<sup>9</sup> which yields high-plex images at subcellular resolution and has been used to characterize immune populations in several tumor types<sup>10–13</sup>. In t-CyCIF, a high-plex image is constructed from a series of 4 to 6 color images, which are then registered and superimposed. The images provide information on the amount of epitope that is expressed as well as the location of the epitope within the tissue. By segmenting the images to demarcate single cells or subcellular compartments, we can then use epitope expression levels to discriminate immune, tumor, and stromal cell types and compute their numbers and distributions within tumors and surrounding normal tissue.

The quality of the antibody reagents largely dictates the reliability of data that is generated by antibody-based imaging methods such as multiplexed ion beam imaging (MIBI)<sup>14</sup>, imaging mass cytometry (IMC)<sup>15</sup>, co-detection by indexing (CODEX)<sup>16</sup>, DNA exchange imaging (DEI)<sup>17</sup>, MultiOmyx (MxIF)<sup>18</sup>, imaging cyler microscopy (ICM)<sup>19–21</sup>, multiplexed IHC<sup>22</sup>, NanoString Digital Spatial Profiling (DSP)<sup>23</sup>, and t-CyCIF itself. We have recently published detailed methods for validating antibodies and assembling panels of antibodies for multiplexed tissue techniques<sup>24</sup>. That work highlights a variety of complementary approaches to qualify antibodies using information at the level of pixels, cells, and tissues and yielded a 16-plex antibody panel capable of detecting lymphocytes, macrophages, and immune checkpoint regulators for use in ‘immune profiling’ tissue samples. Using t-CyCIF, we qualified antibodies in reactive (non-neoplastic) tonsil tissue (TONSIL-1), which has a highly stereotyped arrangement of diverse immune cell types, and then demonstrated the panel’s utility in characterizing common and rare immune populations in three lung cancer tissue specimens: a lung adenocarcinoma that had metastasized to a lymph node (LUNG-1-LN), a lung squamous cell carcinoma that had metastasized to the brain (LUNG-2-BR), and a primary lung squamous cell carcinoma (LUNG-3-PR). We also provide t-CyCIF imaging data from eight FFPE sections used to validate antibodies; in these samples, antibodies were applied in different permutations and order, making the data useful for examining relationships between antigenicity, fluorescence signal, and cycle number.

In this data descriptor, we share the images from our recent work<sup>24</sup>. The dataset includes immunofluorescence images from formalin fixed paraffin embedded (FFPE) tissue sections mounted onto glass slides. In each section, there are between ~61,800 to ~483,000 individual cells with fluorescence intensity and spatial information provided for 27 antibodies that were acquired in a multiplexed fashion. These antibodies include the highly validated 16-plex immune panel as well as antibodies against several additional markers of interest such as markers of tumor cell lineage and cell proliferation. We also include quantitative, single-cell measurements of 60+ features including fluorescence intensity measurements for each target epitope/protein, cellular morphology measurements such as area, eccentricity, and solidity, and spatial information such as the centroid position of each cell and its nearest neighbors.

The resulting single-cell data can be analyzed using qualitative and quantitative approaches both in the context of the original spatial arrangement of the tissue and as sets of derived feature vectors, one for each cell. Spatial views enable the analysis of geographic patterns and interactions between different cell types, such as the immune microenvironment surrounding tumor tissue. Such data can be used to develop new methods for visualizing large complex images and to develop and refine data analysis approaches such as image segmentation, intensity gating (to discriminate ‘positive’ and ‘negative’ cell populations), and spatial clustering. As multiple research centers begin to assemble high-dimensional and multi-parametric atlases of human cancers and pre-cancers<sup>25</sup>, there is an increasing need for cross-center validation of analysis methodologies. Publicly available datasets such as ours will provide a freely accessible resource for such efforts.

## Methods

**Tissue samples.** Five formalin-fixed paraffin-embedded (FFPE) human tissue samples were retrieved from the archives of the Department of Pathology at Brigham and Women’s Hospital with IRB approval as part of a discarded tissue protocol. The diagnoses were confirmed by a board-certified pathologist (S.S.) (Table 1). Sections were cut from FFPE blocks at a thickness of 5  $\mu$ m and mounted onto Superfrost Plus microscope slides prior to use.

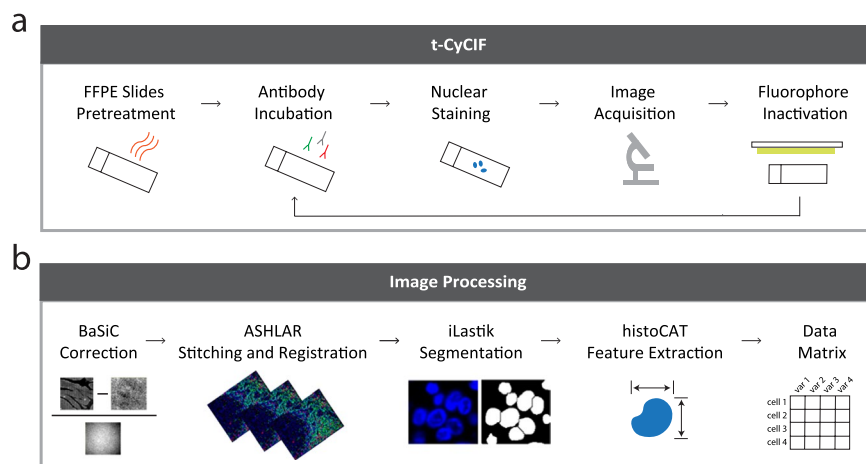
**Datasets.** Data from tissue samples was acquired in two batches. The first batch (DATASET-1) contains data from LUNG-1-LN, LUNG-2-BR, LUNG-3-PR, and TONSIL-1. The second batch (DATASET-2) contains data from eight sections of TONSIL-2. Data associated with each of these sections are labeled TONSIL-2.1, TONSIL-2.2, etc. in the data records. Note that in the sample coding system, the number after the dash denotes patient sample and the number after the decimal point denotes block section.

**Tissue-based cyclic immunofluorescence.** Each section of tissue was imaged with a panel of 26–28 antibodies using t-CyCIF as previously described<sup>9</sup>. This method consists of iterative cycles of antibody incubation, imaging, and fluorophore inactivation (Fig. 1).

**Slide preparation.** An automated program on the Leica Bond RX (Leica Biosystems) was used to prepare slides for t-CyCIF. The slides were treated as follows: baked at 60 °C for 30 min, dewaxed at 72 °C with Bond Dewax Solution (Cat. AR9222, Leica Biosystems), and treated with Epitope Retrieval 1 (ER1) Solution at 100 °C for 20 min for antigen retrieval. Odyssey Blocking Buffer (Cat. 927–40150, LI-COR) was applied to the slides at room temperature (RT) for 30 min and then incubated with three secondary antibodies at RT for 60 min, followed by Hoechst 33342 (Cat. H3570, Life Technologies) solution (2  $\mu$ g/ml) at RT for 30 min.

Sample Code	Data Set	Tissue Type	Clinical Classification
TONSIL-1	1	Human tonsil tissue	Normal tonsil
LUNG-1-LN	1	Human lung carcinoma tissue	Lung adenocarcinoma metastasis to lymph node
LUNG-2-BR	1	Human lung carcinoma tissue	Lung squamous cell carcinoma metastasis to brain
LUNG-3-PR	1	Human lung carcinoma tissue	Primary lung squamous cell carcinoma
TONSIL-2.1	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.2	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.3	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.4	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.5	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.6	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.7	2	Human tonsil tissue	Reactive tonsil
TONSIL-2.8	2	Human tonsil tissue	Reactive tonsil

**Table 1.** Sample Information.



**Fig. 1** Overview of data generation. **(a)** Multiplexed, immunofluorescence images were acquired using the tissue-based cyclic immunofluorescence (t-CyCIF) method and **(b)** processed with a series of algorithms and toolboxes including BaSiC, ASHLAR, ilastik, and histoCAT to obtain single-cell features.

**Blocking.** After slide preparation, non-specific, reactive epitopes were blocked by incubating slides overnight at 4°C in the dark with fluorescently conjugated secondary antibodies raised against the host species of the unconjugated, primary antibodies used in the first cycle of t-CyCIF.

**Antibody staining.** Slides were initially imaged to measure nonspecific binding from secondary antibodies, photobleached, and then imaged again to measure tissue autofluorescence. In the first cycle of antibody incubation, the slides were incubated overnight with primary antibodies from different species and then with corresponding secondary antibodies for two hours at RT in the dark. Slides were then washed with 1X PBS, stained with Hoechst solution, and then imaged. This process was repeated for 11–12 cycles using antibodies directly conjugated to fluorophores. All antibodies used in this study are listed in Online-only Table 1 with an assigned unique identifier. Antibodies and imaging parameters used for each cycle of imaging for all samples in DATASET-1 are detailed in Online-only Table 2 and for all samples in DATASET-2 in Online-only Table 3.

**Mounting and de-coverslipping.** Prior to each cycle of imaging, slides were wet-mounted using 200 µl of 10% glycerol in PBS and 24 × 50 mm glass cover slips (Cat # 48393-081, VWR). Following imaging, the slides were de-coverslipped by placing the slides vertically in a slide rack completely submerged in a container of 1X PBS for 15 minutes and slowly pulling the slides back up, allowing the glass coverslip to remain in the PBS.

**Image acquisition.** Images from each cycle of t-CyCIF were acquired using the RareCyte CyteFinder Slide Scanning Fluorescence Microscope. The four following filter sets were used: 1) The ‘DAPI channel’ for imaging Hoechst with a peak excitation of 390 nm and half-width of 18 nm and a peak emission of 435 nm and half-width of 48 nm, 2) the ‘488 channel’ with a 475-nm/28-nm excitation filter and a 525-nm/48-nm emission filter, 3) the ‘555 channel’ with a 542-nm/27-nm excitation filter and a 597-nm/45-nm emission filter, and 4) the ‘647 channel’ with a 632-nm/22-nm excitation filter and a 679-nm/34-nm emission filter. Each tissue section was imaged twice, a large region with a 10X/0.3 NA objective and a smaller region with a 40X/0.6NA objective. The 10X images

have a field of view of  $1.6 \times 1.4$  mm and a nominal resolution of  $1.06 \mu\text{m}$ . The 40X images have a field of view of  $0.42 \times 0.35$  mm and a nominal resolution of  $0.53 \mu\text{m}$ . For both sets of images, a 5% overlap was collected between fields of view to facilitate image stitching. In DATASET-2, the first cycle of antibodies was imaged twice, once with a high exposure time and once with a low exposure time.

**Photobleaching.** Following slide preparation using the Leica Bond RX and subsequent to each cycle of imaging, fluorophores were inactivated by submerging slides in a solution of 4.5%  $\text{H}_2\text{O}_2$  and 20 mM NaOH in 1X PBS and incubating them under a light emitting diode (LED) for 2 hours at RT.

**Image processing.** *Background and shading correction.* The BaSiC algorithm<sup>26</sup> plugin for ImageJ was used to computationally derive flat-field and dark-field profiles from the original image for each cycle. The flat-field is used to correct for irregular illumination of the sample, and the dark-field is used to correct for camera sensor offset and internal noise. Lambda values of 0.1 and 0.01 were used for flat-field and dark-field, respectively. For each cycle, the raw image was subtracted by the dark-field profile and divided by the flat-field profile to correct the shading on each individual image field.

*Stitching and registration.* ASHLAR (version v1.6.0) was used to stitch the fields from the first imaging cycle into a mosaic and to co-register the fields from successive cycles of imaging. Ashlar stitches fields together by calculating the phase correlation between neighboring images to correct for local state positioning error and applying a statistical model of microscope stage behavior to correct for large-scale error. It then uses a similar phase correlation approach to register fields from successive cycles to the first cycle of stitched images. The output is an OME-TIFF file that contains a seamless multi-channel mosaic depicting the entire sample across all image cycles.

*Segmentation.* The OME-TIFF output from ASHLAR was used to segment single cells in the images using the ilastik software program<sup>27</sup> and MATLAB (version 2018a). The OME-TIFF was cropped into  $6000 \times 6000$  pixel regions to increase processing speed. From each cropped region,  $\sim 20$  random  $250 \times 250$  pixel regions were selected and used as training data in the ilastik program to generate a probability of each pixel in the cropped region belonging to three classes: nuclear area, cytoplasmic area, or area not occupied by a cell (background). During the labeling process, the user was presented with the DAPI channel only. The user labeled pixels with DAPI as nuclei, pixels on the border or a few pixels away from DAPI signal as cytoplasm, and pixels distant from DAPI signal as background. While labeling by the user was performed using only one DAPI channel, all 44 channels from the stitched and registered images were used by ilastik to train the pixel classification algorithm. Color/intensity features including gaussian smoothing, edge features including the Laplacian of gaussian, gaussian of gradient magnitude, and difference of gaussians, and texture features including structure tensor eigenvalues and hessian of gaussian eigenvalues with a  $\sigma_0 = 0.30$ ,  $\sigma_1 = 0.70$ ,  $\sigma_2 = 1.00$ ,  $\sigma_3 = 1.60$ ,  $\sigma_4 = 3.50$ , and  $\sigma_5 = 5.03$  were used to train the pixel classification in ilastik. The ilastik software generated three probability masks, one for each of the three classes. For example, the cytoplasmic probability mask was a TIFF image, with each pixel containing a value between 0 to 65535 where larger values indicate higher probability of that pixel belonging to the cytoplasmic class. The probability masks along with morphological manipulations were used in MATLAB to perform a watershed transformation and identify objects, or cell nuclei. The output from MATLAB was a nuclear segmentation mask for each cropped region. Please see below for a description of the qualitative and quantitative approaches we used for the technical validation and assessment of the segmentation.

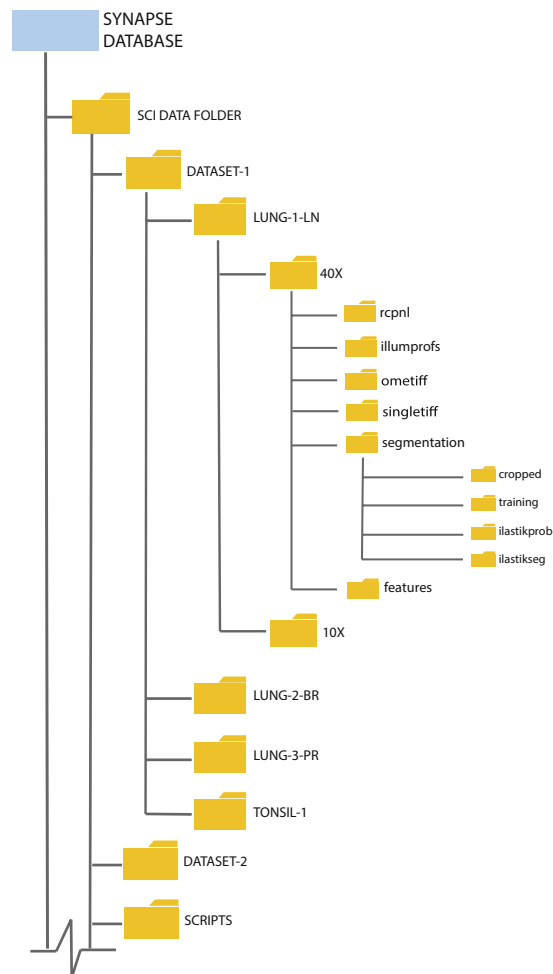
*Single-cell feature extraction.* The histology topography cytometry analysis toolbox (histoCAT)<sup>28</sup> was used to extract features of the cells segmented in each image. Single cell features included fluorescence intensity measurements of each antibody, morphological features such as cell area and circularity, as well as spatial features such as the centroid position of the cell. Moreover, cells in spatial proximity to one another were identified and indexed to enable neighborhood analysis and cell phenotype interactions. The output was a data table for each cropped region. For each sample, the data tables from all the cropped regions were concatenated into a master image level data table with each cell assigned a global unique identifier and centroid position. A complete list and description of each feature in the master data tables is provided in Online-only Table 4.

## Data Records

We have made all the data for this manuscript available in the Synapse repository hosted by Sage Bionetworks <https://doi.org/10.7303/syn17865732><sup>29</sup>. We organized the data as described in Fig. 2. For each tissue sample, we share image data acquired at two magnifications. For each 40X magnification in DATASET-1, we share:

- i. raw rcpln files,
- ii. illumination profiles generated by the BaSiC algorithm,
- iii. an OME-TIFF file output from the ASHLAR algorithm,
- iv. individual TIFF images for each marker,
- v. probability masks for segmentation from ilastik software,
- vi. labeled nuclear segmentation mask, and
- vii. data table of 60+ features extracted for each cell.

The “rcpln” folder contains the raw image files in an rcpln file format generated by the RareCyte CyteFinder for each cycle of imaging. The “illumprof” folder contains TIFF files for the dark-field profile and the flat-field profile for each cycle of imaging. Each TIFF file in this folder is a stack of four TIFF images corresponding to the four wavelengths imaged every cycle. The “ometiff” folder contains one OME-TIFF file that is a stitched, registered



**Fig. 2** Database structure. All shared data are stored in the SYNAPSE repository. <https://doi.org/10.7303/syn17865732>

mosaic of all channels across all cycles of imaging. The OME-TIFF file has a pyramidal structure that contains mosaics at multiple resolutions. The “singletiff” folder contains a single TIFF mosaic for each marker at the highest resolution. This folder separates the OME-TIFF into separate channels to facilitate opening in software that is incompatible with the OME-TIFF format. The “segmentation” folder contains subfolders with intermediate data outputs from the segmentation process. The “cropped” subfolder contains  $6000 \times 6000$  pixel regions from the OME-TIFF file. The “training” subfolder contains  $250 \times 250$  pixel regions used as training data for segmentation. The “ilastikprob” subfolder contains a TIFF image for the probability of each pixel in the cropped regions belonging to each class used in ilastik training. The “ilastikseg” folder contains a TIFF image of the nuclear segmentation mask. This folder also contains an TIFF image stack with the segmentation mask and the DAPI fluorescence image from the first cycle of imaging for easy comparison of the accuracy of the probability mask. The “features” folder contains a csv data table for each cropped region with 60+ feature measurements for each cell as well as a master data table with data from each cropped region combined. Note that the X and Y coordinates for the centroid of the cell in the master table reflects the global position of the cell in the entire piece of tissue imaged/stitched image.

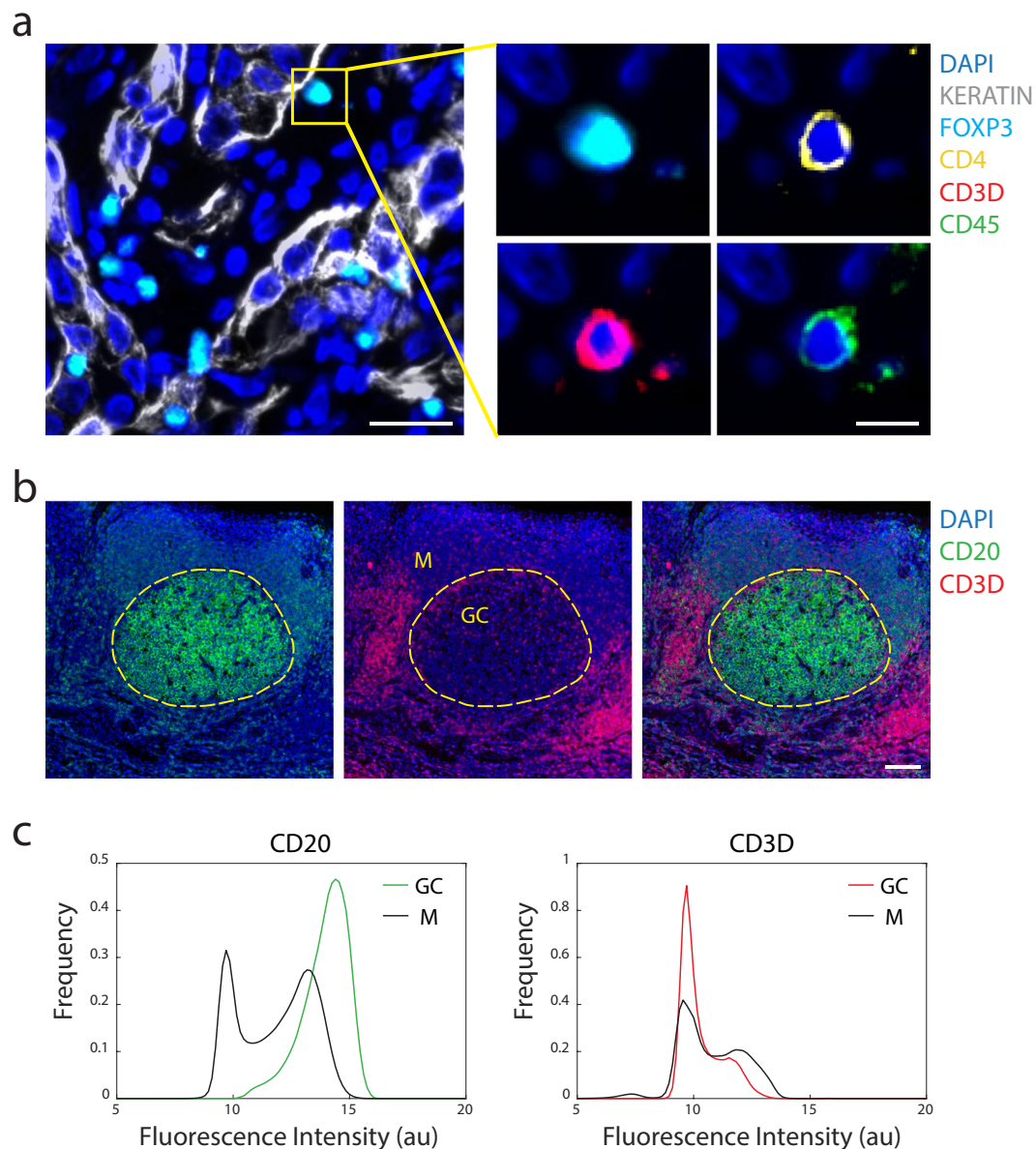
We provide all scripts used in data generation. A description of the scripts and supporting documents is provided in Online-only Table 5.

Additionally, a subset of the imaging data can be found and viewed on cycif.org (<https://www.cycif.org/featured-paper/du-lin-rashid-2019/figures/>). In this interactive image browser, we indicate several distinct regions of interest in the tonsil and lung cancer images and provide descriptive narrations about a subset of the combinations of immune markers expressed in these samples.

### Technical Validation

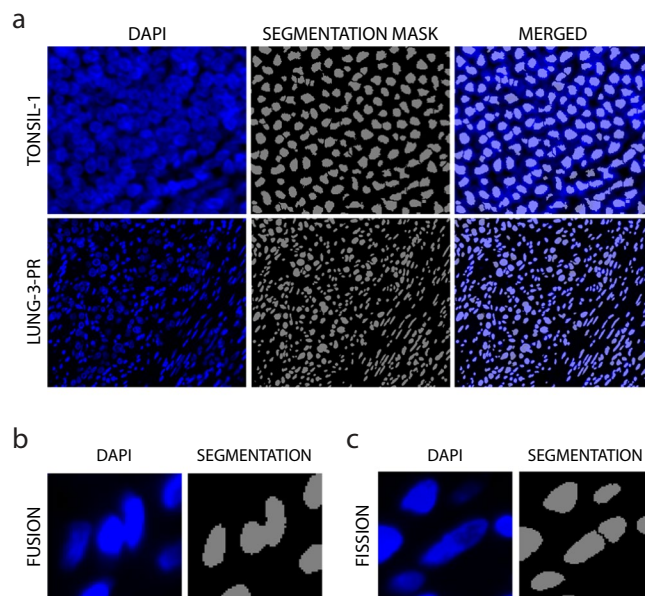
**Staining quality.** We performed a detailed validation of the panel of antibodies used to generate the datasets described in our prior work<sup>24</sup>. One or more trained pathologists visually reviewed the staining patterns for each antibody to assess specificity to cell type, appropriate localization within the cell (e.g. nucleus v. cytoplasm v. membrane), co-staining with other markers, and localization to the expected geographic regions within the tissue. For example, the cytokeratin antibody, known to detect intermediate filament proteins in epithelial cells,





**Fig. 3** Antibody staining quality. **(a)** Immunofluorescence image from LUNG-3-PR showing epithelial tumor cells marked by Keratin (white) and a regulatory T cell marked by FOXP3 (cyan), CD4 (yellow), CD3D (red), and CD45 (green) (scale bar: 25  $\mu$ m; inset scale bar: 10  $\mu$ m). **(b)** A region of TONSIL-1 showing CD20 (green) and CD3D (red) expression. Area inside yellow dashed circle denotes germinal center (GC), and area outside denotes the mantle (M) region (scale bar: 100  $\mu$ m). **(c)** Probability density function of fluorescence signal intensity of every pixel in the germinal center ( $n = 1,446,450$  pixels) and mantle ( $n = 4,369,358$  pixels) for CD20 and CD3D within the region shown in **(b)**. X-axis is fluorescence intensity ( $\log_2$  au) and y-axis is frequency of pixels.

was expressed in striated patterns surrounding the nuclei of cells morphologically consistent with epithelial origin, whereas the FOXP3 antibody, targeting a transcription factor in T cells, was concentrated in the nuclear area of small, round cells morphologically consistent with lymphocytes (Fig. 3a). Antibodies detecting cell lineage markers such as FOXP3, which delineates a regulatory T-cell population, were further corroborated by assessing appropriate co-expression of other markers. For example, we found that FOXP3 was co-expressed with CD4, CD3D, and CD45, thereby increasing our confidence in the staining quality (Fig. 3a). As another example, CD20, a B-cell antigen, was observed to have higher levels of signal within germinal centers of tonsil tissue which are well-established B cell rich compartments within tonsil rather than the mantle region where we found an abundance of cells expressing the T-cell antigen CD3D (Fig. 3b). See our prior publication<sup>23</sup> for additional quality measurements including the comparison of t-CyCIF antibody staining to the staining observed with clinical grade antibodies that were used in immunohistochemistry (IHC) staining, pixel-by-pixel correlations of multiple antibody clones against the same target, and various high-dimensional cell clustering methods.



**Fig. 4** Assessment of segmentation. (a) Representative images of DAPI staining and corresponding segmentation mask in TONSIL-1 and LUNG-3-PR. (b) Examples of fusion (under-segmentation) and (c) fission/splitting (over-segmentation).

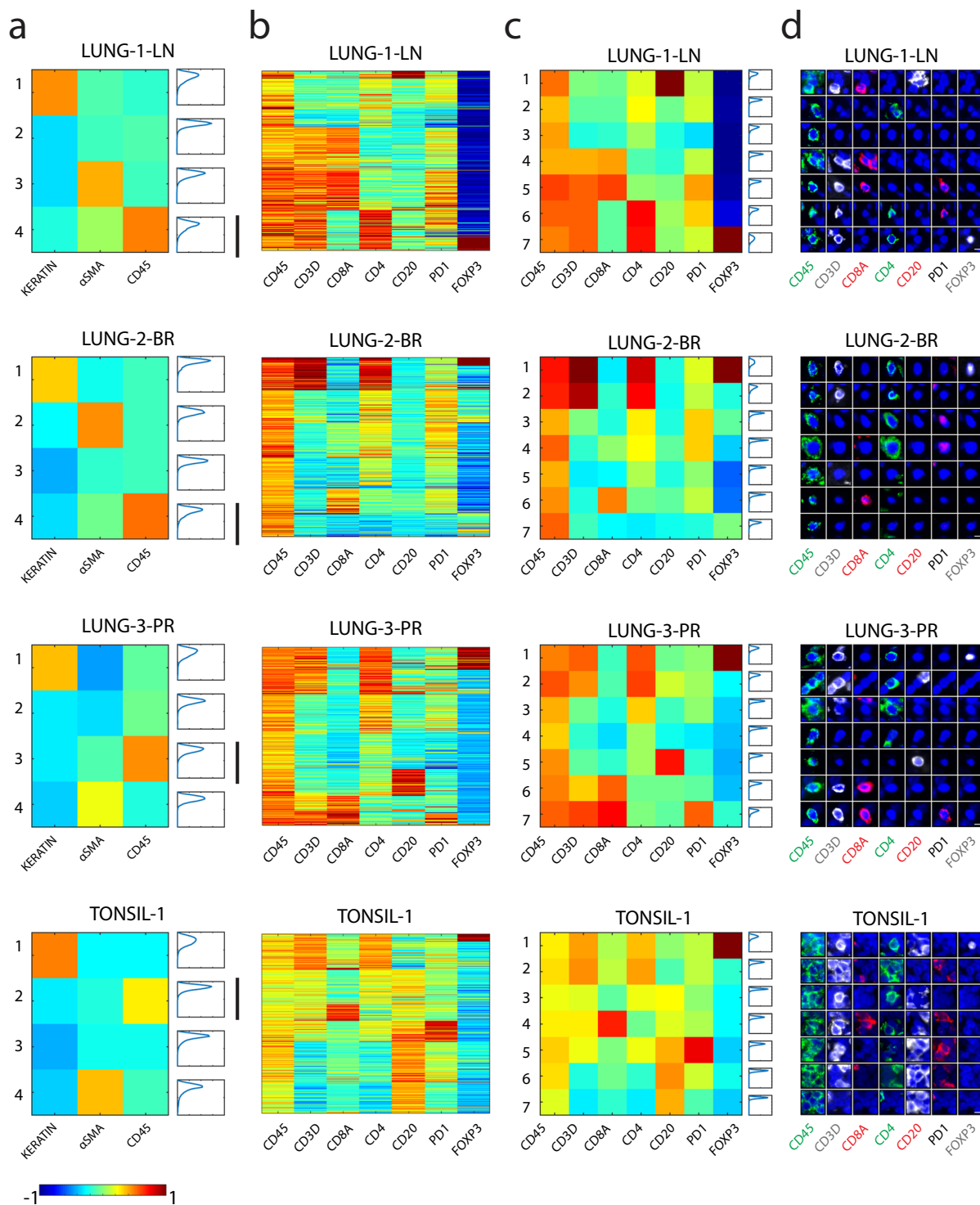
	LUNG-1-LN		LUNG-2-BR		LUNG-3-PR		TONSIL-1	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
True Positive	0.872	0.013	0.890	0.017	0.899	0.013	0.863	0.012
Fusion	0.086	0.027	0.074	0.017	0.053	0.019	0.038	0.020
Fission	0.042	0.017	0.036	0.003	0.047	0.008	0.099	0.025

**Table 2.** Segmentation Accuracy.

**Cell segmentation.** We evaluated the quality of segmentation of single cells within the tissue images using a two-step system. We only performed segmentation on the 40X magnification images because the lower resolution of the 10X magnification images reduced segmentation accuracy. First, we overlaid the segmentation masks over the DAPI signal to evaluate the accuracy of segmentation qualitatively (Fig. 4a); based on these data, we then adjusted and optimized the segmentation. Second, three users evaluated a random sample of 500 cells from the tonsil and each of the lung tissues to quantify the accuracy, or true positives, and rate of fusion errors (under-segmentation) and fission/splitting errors (over-segmentation) among mis-segmented cells (Fig. 4b-c, Table 2). The cell segmentation of all samples had a low error rate (~0.1) across cells of various morphologies (large tumor cells, smaller round immune cells, elongated fibroblasts, etc.). The accuracy of image segmentation can be further improved with the development of new algorithms.

In our analysis of these images, we observed that the area covered by the nuclear mask effectively captured the signal from the nuclear compartment as well as the cytoplasmic/membranous compartment as can be observed in Fig. 3a. The presence of cytoplasmic signal in the nuclear compartment in this dataset is in part attributable to the three-dimensional nature of the five-micron thick tissue sections which we imaged. These sections capture the complex intermingling of nuclear and cytoplasmic compartments that occurs in individual cells. Thus, the signal that is ultimately projected into a two-dimensional image does not arise strictly from one cellular compartment. Moreover, the high cellular density in these tumor and tonsil tissues in combination with high intensity fluorescence signal created conditions where expanding the nuclear segmentation mask captured signal from neighboring cells. Therefore, in our single-cell analyses, we used the nuclear segmentation mask to extract signal intensity features for both nuclear and cytoplasmic markers.

**Single-cell feature extraction.** To assess the integrity of the single-cell features extracted from the images, we applied an unsupervised, k-means clustering method to the data from the three lung cancer resection samples and the reactive (non-neoplastic) tonsil sample. This analysis yielded four cardinal cell types (clusters) using three lineage markers (Fig. 5a). For each sample, the cells clustered into an epithelial group marked by keratin expression, a stromal group marked by  $\alpha$ SMA expression, and an immune group marked by CD45 expression. A fourth group was marked by low expression of all three markers. We then isolated the cells in the immune group and further clustered them using other lymphocyte markers (Fig. 5b,c). The clustering revealed similar immune cell populations to those observed by visual review of the images and as quantified previously using other computational



**Fig. 5** Heatmaps of cell populations from lung cancer and tonsil tissues using k-means clustering demonstrates distinct cell immune populations with expected patterns of biomarker expression. **(a)** Heatmap of the expression of Keratin,  $\alpha$ SMA, and CD45 in all cells that were collected from LUNG-1-LN, LUNG-2-BR, LUNG-3-PR, and TONSIL-1 using k-means clustering. Each row is a cluster. The last column in each heat map shows the probability density function (pdf) plot showing the fit of each cell within the cluster, with the x-axis denoting the frequency of cells and y-axis denoting the Euclidean distance of the cell from the centroid of the cluster. The black vertical bars mark the immune cluster with high CD45 expression. **(b,c)** Heatmaps showing the expression of seven lymphocyte markers (CD45, CD3D, CD8A, CD4, CD20, PD1, FOXP3) from the cells within the CD45 high cluster from panel (a). **(b)** Each row represents protein marker expression data from a single cell or **(c)** each row represents a cluster. Note that fluorescence intensity values were log transformed and normalized between  $-1$  to  $1$  as indicated by the color bar. **(d)** Galleries of immunofluorescence images of representative cells from each cluster in (c). (Scale bar:  $5\ \mu\text{m}$ ).



methods<sup>24</sup>. Each cluster exhibited varying degrees of tightness, or fit. The probability density function plot for each cluster in Fig. 5a,c displays the distance of each cell from the centroid of the cluster, with the y-axis denoting distance and the x-axis denoting the frequency of cells belonging to each distance bin. The range of the curve along the y-axis reflects the fit, with a smaller range denoting greater fit and a larger range denoting poorer fit. The variability of cluster fit can be explained by the intrinsic heterogeneity within different immune populations. Tighter clusters where the majority of the cells have short distances from the center represent populations with distinct and highly similar marker expression profiles. Looser clusters, with wider distance ranges and longer tails, likely contain subpopulations of immune cells that may require further stratification and investigation. While this exercise displays fundamental immune cell populations reported in the literature, we note the potential of multiplexed data and unsupervised methods to reveal novel cell populations and states. Here, using alternative segmentation, feature extraction, and computational approaches, we retained reproducible immune cell populations, giving us confidence in the robustness of this dataset.

## Usage Notes

More information on the t-CyCIF method used to generate this data can be found at: [www.cycif.org](http://www.cycif.org) and a detailed protocol can be found in Lin *et al.*<sup>9</sup> and Du, Lin, Rashid *et al.* 2019<sup>24</sup>.

A narrative of the dataset is available for interactive web-browsing here: <https://www.cycif.org/featured-paper/du-lin-rashid-2019/figures/>.

Open data agreement: Highly multiplexed immunofluorescence images and single-cell data of immune markers in tonsil and lung cancer by Rumana Rashid, Giorgio Gaglia, Yu-An Chen, Jia-Ren Lin, Ziming Du, Zoltan Maliga, Denis Schapiro, Clarence Yapp, Jeremy Muhlich, Artem Sokolov, Peter Sorger and Sandro Santagata is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Based on a work at <https://doi.org/10.7303/syn17865732>.

## Code availability

All code used to process and generate the data in this study can be found alongside the dataset<sup>29</sup>. A description of each script is provided in Online-only Table 5. Source code for ASHLAR is available on GitHub (<https://github.com/jmuhlich/ashlar>). The newest histoCAT version can also be found GitHub (<https://github.com/BodenmillerGroup/histoCAT>).

Received: 20 June 2019; Accepted: 21 November 2019;

Published online: 17 December 2019

## References

1. Wei, S. C., Duffy, C. R. & Allison, J. P. Fundamental Mechanisms of Immune Checkpoint Blockade Therapy. *Cancer Discov.* **8**, 1069–1086 (2018).
2. Sharma, P. & Allison, J. P. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell* **161**, 205–214 (2015).
3. Socinski, M. A. *et al.* Atezolizumab for First-Line Treatment of Metastatic Nonsquamous NSCLC. *N. Engl. J. Med.* **378**, 2288–2301 (2018).
4. Carbone, D. P. *et al.* First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2415–2426 (2017).
5. Forde, P. M. *et al.* Neoadjuvant PD-1 Blockade in Resectable Lung Cancer. *N. Engl. J. Med.* **378**, 1976–1986 (2018).
6. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
7. Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321–330 (2017).
8. Bodenmiller, B. Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. *Cell Syst.* **2**, 225–238 (2016).
9. Lin, J.-R. *et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife*, <https://doi.org/10.7554/eLife.31657> (2018).
10. Coy, S. *et al.* Multiplexed immunofluorescence reveals potential PD-1/PD-L1 pathway vulnerabilities in craniopharyngioma. *Neuro-oncology* **20**, 1101–1112 (2018).
11. Dunn, I. F. *et al.* Mismatch repair deficiency in high-grade meningioma: a rare but recurrent event associated with dramatic immune activation and clinical response to PD-1 blockade. *JCO Precis. Oncol.* **2018** (2018).
12. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984–997.e24 (2018).
13. Baker, G. J. *et al.* Systemic Lymphoid Architecture Response Assessment (SYLARAS): An approach to multi-organ, discovery-based immunophenotyping implicates a role for CD45R/B220<sup>+</sup> CD8T cells in glioblastoma immunology. Preprint at, <https://doi.org/10.1101/555854> (2019).
14. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).
15. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
16. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968–981.e15 (2018).
17. Wang, Y. *et al.* Rapid Sequential *in Situ* Multiplexing with DNA Exchange Imaging in Neuronal Cells and Tissues. *Nano Lett.* **17**, 6131–6139 (2017).
18. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. USA* **110**, 11982–11987 (2013).
19. Schubert, W. *et al.* Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat. Biotechnol.* **24**, 1270–1278 (2006).
20. Friedenberger, M., Bode, M., Krusche, A. & Schubert, W. Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome imaging system: sample preparation and measuring procedures. *Nat. Protoc.* **2**, 2285–2294 (2007).
21. Hillert, R. *et al.* Large molecular systems landscape uncovers T cell trapping in human skin cancer. *Sci. Rep.* **6**, 19012–19012 (2016).
22. Tsujikawa, T. *et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Reports* **19**, 203–217 (2017).
23. Decalf, J., Albert, M. L. & Ziai, J. New tools for pathology: a user's review of a highly multiplexed method for *in situ* analysis of protein and RNA expression in tissue. *J. Pathol.* **247**, 650–661 (2019).

24. Du, Z. *et al.* Qualifying antibodies for image-based immune profiling and multiplexed tissue imaging. - PubMed - NCBI. *Nature Protocols* **14**, 2900–2930 (2019).
25. Srivastava, S. *et al.* The Making of a PreCancer Atlas: Promises, Challenges, and Opportunities. *Trends in Cancer* **4**, 523–536 (2018).
26. Peng, T. *et al.* A BaSiC tool for background and shading correction of optical microscopy images. *Nat. Commun.* **8**, 14836 (2017).
27. Sommer, C., Strähle, C., Köthe, U. & Hamprecht, F. ilastik: Interactive Learning and Segmentation Toolkit. In *Chicago. Proceedings* 230–233.
28. Schapiro, D. *et al.* histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods* **14**, 873–876 (2017).
29. Rashid, R. *et al.* Highly multiplexed immunofluorescence images and single-cell data of immune markers in tonsil and lung cancer. *Synapse*, <https://doi.org/10.7303/syn17865732> (2019).

### Acknowledgements

This work was funded by NIH grants U54-CA225088 and U2C-CA233262 to P.K.S. and S.S., by U2C-CA233280 to P.K.S., and by the Ludwig Center at Harvard. The Dana-Farber/Harvard Cancer Center is supported in part by an NCI Cancer Center Support Grant P30-CA06516. D.S. was supported by the BioEntrepreneur-Fellowship of the University of Zurich (BIOEF-17-001) and an Early Postdoc Mobility fellowship (P2ZHP3\_181475). G.G. was supported by T32-HL007627.

### Author contributions

R.R., G.G., Y.A.C., J.R.L., Z.D., Z.M., C.Y., D.S., J.M. and A.S. contributed to data collection, processing, and analysis. R.R., P.K.S. and S.S. wrote the manuscript. S.S. and P.K.S. supervised the project.

### Competing interests

P.K.S. is on SAB of RareCyte, Inc., whose product was used to acquire this data, and Glencoe Software, Inc., whose product was used to visualize this data. S.S. is a consultant for RareCyte, Inc.

### Additional information

**Correspondence** and requests for materials should be addressed to P.S. or S.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019